Date	3rd October 2022
Team ID	PNT2022TMID47507
Project Name	Web Phishing Detection
Team Leader	Gayathri.A
Team Members	Santhiya.S, Shally Therse.P, Suruthi.S, Muthupriya.G

WEB PHISHING DETECTION

ABSTRACT:

In today's era, due to the surge in the usage of the internet and other online platforms, security has been major attention. Many cyberattacks take place each day out of which website phishing is the most common issue. It is an act of imitating a legitimate website and thereby tricking the users and stealing their sensitive information. There are e-banking websites that ask users to provide sensitive data such as username, password & credit card details, etc.., often for malicious reasons. This type of e-banking website is known as a phishing website. So, concerning this problem, this will introduce a possible solution to avoid such attacks by checking whether the provided URLs are phishing URLs or legitimate URLs. It is a Machine Learning based system especially Supervised learning where we have provided many of phishing and legitimate URL dataset. We have taken into consideration the Random Forest Algorithm due to its performance and accuracy. It considers such features and hence detects whether the URL is safe to access or a phishing URL.

LITERATURE SURVEY

- In this survey paper they have mentioned how phishing attacks appear, how the phishers use email or message, as evidence to target the individual or business by sending the link to victim people and deceive them with a large no of phishing emails or messages every day, so many of the corporations or individual are not able to recognize them all. So, here they have mentioned various types of phishing attacks like Learning Model Algorithm, Naive Bayes Algorithm, Decision tree, SVM (Support Vector Machine), Artificial Neural Network and many more. They have also mention phishing detection approaches like the Heuristic-based Approach, Fuzzy-based Approach, Machine Learning Approach, Image-based Approach, and so on.

- This approach makes use of the Naïve Bayes, SMO, J48 algorithm are used for feature selection. There are several separate processes. The first process is to extract the properties of URLs and generate a matrix then secondary process uses the attribute-based feature selection technique to specify the prominent properties after using the attribute-based technique, the new dataset is used as input data to the Machine Learning Algorithm to analyse the website is legitimate or not. Based on the classification method on J48(Decision Tree), Naive Bayes, and SMO (Sequential Minimal Optimization). Here SMO & J48 shows their best accuracy output result was as Naive Bayes performed poorly and it is the least recommended method among all the methods.

- This paper illustrates various types of Phishing Techniques and Anti-Phishing technique because phishing attack is one of a version of harmful content which has found recently a wide circulation in an information field of the modern switched communication systems. So, to identify the website is legitimate or not so there is some feature through which we can identify that the website is legitimate or not. If we enter the URL first it will be checked in the blacklist or whitelist if it is a blacklist that means it is a phishing URL else it is a legitimate URL (whitelist)

- The gap mentioned in "Phishing Websites Detection using Machine Learning" is that they have used a small dataset of 1300 URLs where we overcome this by using 4000 datasets from phishtank.com.

- "Phishing Website Detection based on Machine Learning: A Survey" is a survey paper that discusses different types of attacks and anti-phishing approaches. Also, some defence techniques for phishing are mentioned.

S.no	Paper Title	Techniques	Publish Year	Limitations
1	OFS-NN: "An Effective Phishing Websites Detection Model Based on Optimal Feature Selection and Neural Network"	Proposed method has 3 stages:1. Defines a new index -FVV. 2. Designs an optimal feature selection algorithm.3. Produce the OFS-NN model	2019	The continuous growing of features that are sensitive of phishing attacks need collection of more features for the OFS
2	"Fuzzy Rough Set Feature Selection to Enhance Phishing Attack Detection"	The proposed method uses Fuzzy Rough Set (FRS) theory to identify the features. The decision boundary is decided lower and upper approximation region. Using the lower and upper approximation memberships, a set member is decided to which category it belongs	2019	The specific features used in the method is not specified.
3	"Phishing Detection in Websites using Parse Tree Validation"	If the number of recurrence of root node is: 1. more than half the number of nodes, then probability of authenticity is more. 2.quarter the number of nodes, the probability of authenticity is moderate. 3. less than the quarter number of nodes, then probability of authenticity is low which means phishing probability is high.	2018	The false negative and false positive rates are high.

4	"A new method for Detection of Phishing Websites: URL-Detection"	The three major phases in this work are Parsing, Heuristic Classification of data, Performance Analysis in this model. All of these phases use various and distinctive methods for data processing to get results that are better.	2018	Does not give full information about the techniques used.
5	"Phish Box: An approach for phishing validation and detection"	The approach that is pro-posed makes use of 2 phase detection model to increase its performance. 1. An ensemble model is designed for validating the phishing data and for decreasing the cost of labelling manually Active learning is applied. 2.The model for detection is being trained using these validated data.	2018	The black-list contained invalid data when monitored with an interval set as 12 hours
6	"Fresh Phish: A framework for Auto-Detection of Phishing Websites"	This framework was developed considering there are no other open source frame works which, for a given website, measures the features.	2017	Less accuracy and assumption of the dataset considered for legitimate website is accurate.
7	"Phishing Sites Detection based on C4.5 Decision Tree Algorithm"	The approach proposed makes use of features that were extracted from the URL to make decision about the legitimacy of the URL given as input. To generate the rules, the c4.5 algorithm was used. The rules produced are utilized to order the submitted URL as genuine or phishing with better productivity.	2017	Overall accuracy is less as the paper con-siders limited URL features.